Analysis on Romance Movies with the IMDb Dataset

1. Introduction

In this project, we use the IMDb datasets which consist of information about video resources across the world that are documented by the IMDb website. Our project mainly focuses on analyzing the past, the present and the future of romance movies, trying to derive some internal pattern of the romance movie industry.

There are three parts in our project. First, we complete a movie genre trend analysis in general. We aim to visualize and analyze movie trend in the past century, categorized by genres. We start with romance movies, and later expand the scope to four genres: romance, action, science fiction and horror. We mainly focus on how people's taste of movie genres changes over time in this section.

Second, we complete a romance movie export analysis. We want to understand which countries export most romance movies and to which countries are exported. The objection is to visualize with an interactive map which could indicate the number of exports and relationships between countries in terms of romance movies.

Finally, we utilize machine learning to find an optimal classifier that classifies romance movies into two categories: recommended and unrecommended. After training, the classifier can tell the user whether a romance movie deserves watching or not given some basic information of the movie as input.

2. Dataset Introduction

According to the documentation on the IMDb website, each dataset is contained in a gzipped tsv file in the UTF-8 character set. In each file, the first line contains headers that describe the content of each column. Moreover, a *'\N'* is used to denote that a particular field is missing.

The dataset the specific features we used are:

title.akas.tsv.gz :

- tconst (string) - an alphanumeric unique identifier of the title
- region (string) - the region for this version of the title
- language (string) - the language of the title
- types (array) - Enumerated set of attributes for this alternative title

title.basics.tsv.gz

- tconst (string) - alphanumeric unique identifier of the title
- titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- originalTitle (string) - original title, in the original language
- startYear (YYYY) – represents the release year of a title
- genres (string array) – includes up to three genres associated with the title

title.ratings.tsv.gz

- tconst (string) - alphanumeric unique identifier of the title
- averageRating – weighted average of all the individual user ratings
- numVotes - number of votes the title has received

Besides, we used another online resource.

Country_loc.tsv

It includes the country name (str), Alpha-2 code (str), latitude (average) (float) and longitude (average) (float). This dataset mainly serves to provide geographic information to help plot our findings on a world map.

3. Team Member Introduction

Sihan Min is the main contributor to part I of this project. She mainly focused on initial data cleaning. She also visualized and analyzed the movie trends over the last century, categorized by genres.

Jinghui Song is the main contributor to part II of this project. He processed and reorganized data from different datasets. He also geographically visualized the exports of romance movies with plotly package. He also helped reorganizing the team report.

Junhong Shen is the main contributor to part III of this project. She attempted to train multiple classifiers that predict a romance movie's quality based on input features such as types, region and director. In the process, she tries to analyze the properties of each classifier associated with this specific dataset. She is also responsible for making the powerpoint and reorganizing the team report.